2005

Data Warehouse (DWH) Efficiency Evaluation Using Fuzzy Data Envelopment Analysis (FDEA)

Lei Zhang
University of Colorado at Denver, lei_zhang_00@yahoo.com

Michael Mannino
University of Colorado at Denver, michael.mannino@cudenver.edu

Diswadip Ghosh
University of Colorado at Denver, bghosh@yahoo.com

Judy Scott
University of Colorado at Denver, judy.scott@ucdenver.edu

Data Warehouse (DWH) Efficiency Evaluation Using Fuzzy Data Envelopment Analysis (FDEA)

Lei Zhang
Business School, University of Colorado at Denver
Lei_zhang_00@yahoo.com

Michael Mannino
Business School, University of Colorado at Denver
Michael.Mannino@cudenver.edu

Biswadip Ghosh
Business School, University of Colorado at Denver
bghosh@yahoo.com

Judy Scott
Business School, University of Colorado at Denver
judy.scott@cudenver.edu

ABSTRACT

This paper proposes the macro model and micro model for the efficiency evaluation of data warehouses (DWH), and introduces Fuzzy Data Envelopment Analysis (FDEA) and Data Envelopment Analysis (DEA) to comprehensively assess DWH efficiency. FDEA can determine the efficiencies of data warehouses with multiple input / output variables, and also resolve the problem that standard DEA cannot handle with imprecise input / output data. FDEA results present efficiency from different viewpoints: Best-Worst Scenario and Worst-Best Scenario. The final ranking results between FDEA and DEA are similar on the highest efficiency and the lowest efficiency of data warehouses. Compared with DEA, FDEA can distinguish the efficiency difference for some data warehouses.

Keywords

Fuzzy Data Envelopment Analysis (FDEA), Data Envelopment Analysis (DEA), Data Warehouse (DWH), Efficiency Model, Decision Making Unit (DMU).

INTRODUCTION

From the middle 1990s, data warehouse (DWH) has become one of the most important tools for information storage and exchange in IT fields. The relevant IS research papers also came out since then. Most of those papers focus on data quality of DWH (Orr 1998; Pipno et al 2002; Jarke et al. 2000). Some papers focus on data refreshment (Bouzeghoub et al. 1999; Jarke et al. 2000; Mannino and Walter 2003). There are few papers involving success factors affecting DWHs (Wixom and Watson 2001; Shin 2003).

Those research papers normally put emphasis on some certain factors influencing DWH implementation. There is no quantitative research to measure the efficiency of DWH from its input and output point of view.

In this paper, we present models to evaluate the relative efficiency of operating a DWH. The macro model assesses the efficiency of operating all DWHs in an organization for both refresh and query production. The macro model consists of an initial stage relating resources (labor and computing) to intermediate output measures (logical and physical size) and a second stage relating the intermediate output variables to measures of system usage (number of queries, number of active users, and connect time). The micro models assess the efficiency of operating an individual DWH for either data refresh or query production. Both micro models relate resources to measures of data quality and system usage. To assess the models, we analyze the refresh efficiency of a hypothetical set of DWHs using three variations of Data Envelopment Analysis (constant returns to scale, variable returns to scale, and Fuzzy DEA). The analysis indicates the more pessimistic ratings of Fuzzy DEA and the importance of the number of data sources in efficient refresh operations.

The results of this study have important implications for quantitative evaluation of the efficiency of IT service organizations, particularly those providing complex data products. Efficiency models for complex data products support evaluation of tradeoffs between costs and data quality levels. As far as we are aware, this paper proposes the first efficiency models for DWH operations along with analysis of a preliminary data set. DEA has been used to study information technology investment efficiency (Banker et al., 1990; Lin and Shao, 2000; Shafer and Byrd, 2000; Shao and Shu, 2003; Wang et al., 1997) and software production efficiency (Paradi et al. 1997) but not information technology operations.

RESEARCH GOAL

The goals of this research are:

1. To develop a framework of DWH efficiency model;
2. To interpret of the evaluation results by use of both Fuzzy DEA and standard DEA;
3. To categorize the DWHs based on their efficiency.

DATA WAREHOUSE EFFICIENCY MODELS

To measure the efficiency of DWH operations, we propose models at two levels of detail (Mannino et al. 2004). The macro model assesses the efficiency of an entire DWH service including all DWHs. Although the ideal architecture involves a single DWH, many organizations operate a small number of DWHs rather than one large DWH. In a previous field study (Mannino and Walter 2003), all of the 13 organizations interviewed in the study had a multiple DWHs. Thus, an important element of the macro model is to assess the efficiency of operating all DWHs, rather than a single DWH. In addition, the macro model variables involve a longer time horizon (monthly) than the micro models to capture end-of-period processing.

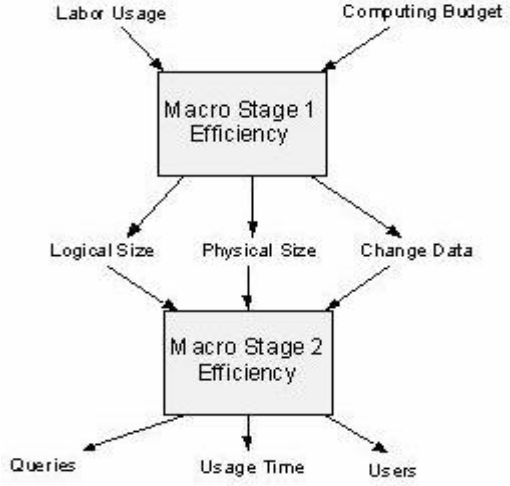


Figure 1: Macro Model of DWH Efficiency

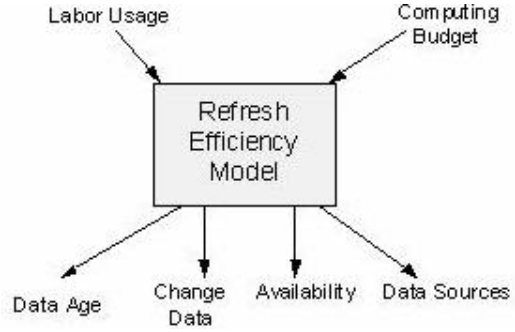


Figure 2: Micro Model of Refresh Efficiency

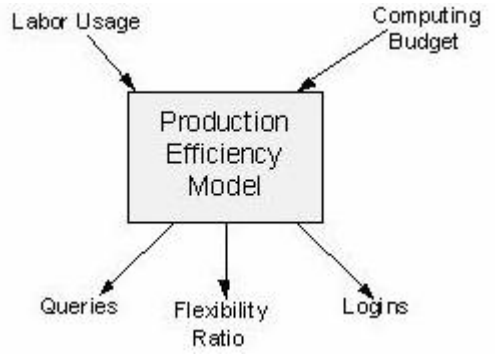


Figure 3: Micro Model of Query Production Efficiency

Variable (Usage)	Meaning	Units of Measure
Labor usage (input)	Labor to support refresh, query production, and help desk activities for all warehouses	Monthly budget ($), Monthly full-time equivalent (FTE) labor hours
Computing budget (input)	Sum of software, hardware, and communication budgets to support operation of all warehouses	Monthly budget ($)
Logical size (output-input)	Indicates the scope of the DWH; Exclude tables used only to facilitate the refresh process	Number of tables
Physical size (output-input)	Total storage size including DWHs, data cubes, materialized views, data marts, and operational data stores	GB
Change data (output-input)	Amount of change data as extracted from data sources before transformation	GB per month or thousands of rows per month
Queries (output)	Number of data requests either directly through ad hoc queries or indirectly through execution of planned reports	Queries per month
Usage time (output)	Total user connect time to all warehouses	Hours per month
Users (output)	Users who login to a DWH site at least once per month	Number of users per month

Table 1: Explanation of Macro Model Variables

Variable (Usage)	Meaning	Units of Measure
Labor usage (input)	Labor to support daily refresh processing for the DWH	Daily budget ($), Daily full-time equivalent (FTE) labor hours
Computing budget (input)	Sum of software, hardware, and communication budgets to support daily refresh processing for the DWH	Daily budget ($)
Data age (output)	Indicates the daily refresh interval for the DWH.	Weighted daily refresh interval in hours
Change data (output)	Amount of change data as extracted from data sources before transformation	GB per day or thousands of rows per day
Availability (output)	Hours of service for user queries; A weighted measure should be used if parts of the warehouse have different availabilities.	Hours per day
Data sources (output)	An operational database or external data owned and maintained outside the organization; The collection of tables in a database is one data source, not separate data sources.	Number of data sources per day

Table 2: Explanation of Refresh Efficiency Model Variables

Variable (Usage)	Meaning	Units of Measure
Labor usage (input)	Labor to support daily query production for the DWH	Daily budget ($), Daily full-time equivalent (FTE) labor hours
Computing budget (input)	Sum of software, hardware, and communication budgets to support daily query production for the DWH	Daily budget ($)
Queries (output)	Number of data requests either directly through ad hoc queries or indirectly through execution of planned reports.	Number of queries per day
Flexibility ratio (output)	Indicates the relative number of ad hoc queries to scheduled queries	Ratio of unplanned to planned queries per day
Usage time (output)	Total user connect time to all warehouses	Hours of connect time per day

Table 3: Explanation of Query Production Model Variables

The macro model consists of multiple stages as depicted in Figure 1 with more detailed explanations of the variables in Table 1. A multiple stage model provides flexibility to evaluate the efficiency of intermediate and final outputs as discussed in (Wang et al., 1997). The inputs (labor usage and computing budget) are the resources used in operating DWHs,

not new development effort to expand the size and scope of a warehouse. Labor usage is specified in both cost and hours to allow comparisons in countries with varying labor rates. Logical size indicates the scope of the DWH, physical size indicates the amount of data in the warehouses, and change data indicates the amount of data extracted from data sources used in the refresh process. Change data comprises new source data (insertions) as well as changed source data (updates and deletions). The final outputs are measures of system usage commonly employed in the information systems success literature (Delone and McClean, 2003).

To allow for imprecision in estimates, the variables can be given as fuzzy numbers rather than point values. A fuzzy number involves an interval and a membership function. The imprecision may involve monthly fluctuations as well as incomplete knowledge. For example, change data can be given as 200 GB to 230 GB to reflect monthly fluctuations. In contrast, usage time may be 5,000 to 6,000 hours to reflect incomplete knowledge about connect time.

The micro models assess the efficiency of operating a single DWH for either daily refresh processing or query production. The emphasis on daily processing fits with the refresh timing in many DWHs found in (Mannino and Walter, 2003) in which organizations invested heavily to provide daily refresh and query production. Since daily operations can vary among DWHs, the micro models involve the efficiency of individual warehouses rather than an organization's entire collection of warehouses. As in the macro model, fuzzy numbers can be used instead of point values to indicate imprecision in the estimates.

The refresh efficiency model consists of a single stage as shown in Figure 2 with variables explained in Table 2. The inputs (labor usage and computing budget) are identical to the macro model except that they are daily numbers limited to refresh processing only. The outputs are different from the macro model except for the change data variable. Data age is a measure of data staleness for a DWH reflecting the time lag between the time an event is recorded in an operational database and the time it is available for DWH users. If there are multiple refresh processes, a weighted data age should be computed reflecting the data age of each batch of change data. In the DEA model, the inverse of data age is used instead of the raw data age because DEA models maximize outputs. Availability is used rather than the time to complete refresh processing because longer refresh processing does not necessarily imply unavailability. With specialized hardware architectures, refresh processing can occur while the warehouse (or at least part of the warehouse) is available for user queries.

The query production efficiency model consists of a single stage as shown in Figure 3 with variables explained in Table 3. The inputs (labor usage and computing budget) are identical to the refresh efficiency model except they are limited to query production only. The queries and usage time variables are identical to the macro model variables. The flexibility ratio variable indicates the extent to which ad hoc queries are supported and used. Planned queries are easier to accommodate than ad hoc queries.

FUZZY DATA ENVELOPMENT ANALYSIS (FDEA)

Data Envelopment Analysis (DEA)

DEA proposed by Charnes and Rhodes 1978 is also called Frontier Analysis. It is a linear programming based technique for measuring the relative efficiency of Decision Making Units (DMU) where the presence of multiple inputs and outputs makes comparisons difficult.

The measurement of relative efficiency where there are multiple possibly incommensurate inputs and outputs can be expressed as the following formula for each Decision Making Unit (DMU) (Charnes and Rhodes 1978). The weights of input and output variables will be computed when the relative efficiency of DMU reaches the maximum value.

$$E_r = \max\left(\frac{\sum_{i=1}^{m} u_i y_{ri}}{\sum_{i=1}^{n} v_i x_{ri}}\right) \quad \text{Subject to:} \quad \frac{\sum_{i=1}^{m} u_i y_{ri}}{\sum_{i=1}^{n} v_i x_{ri}} \le 1 \quad \text{(For all DMUs, } r=1, 2, \ldots, s\text{)}$$

Where E_r is efficiency of unit r; y_{ri} is amount of output i from unit r; x_{ri} is amount of input i from unit r;

u_i is the weight of output i; v_i is the weight of input i; s is the number of DMU;

m is the number of output variables; n is the number of input variables.

DEA models can be of two possible orientations: output and input. Output-Oriented DEA models attempt to achieve efficiency by maximizing the output for a given input, while Input-Oriented DEA models involve minimizing inputs to produce a given output. The selection of output-oriented models or input-oriented models is based on what decision makers

care about: if they have fixed resources, they may choose an output-oriented model; otherwise, if they want to reduce the resource usage, they may choose an input-oriented model.

Returns to scale refers to increasing or decreasing efficiency based on size. Constant Returns to Scale (CRS) means that the producers are able to linearly scale the inputs and outputs without increasing or decreasing efficiency. Variable Returns to Scale (VRS) occurs when the propositional increase in all inputs does not result in the same propositional increase in outputs. If there is a difference of the efficiency score between CRS and VRS for a DMU, it indicates that this DMU has scale inefficiency.

DEA can evaluate the relative efficiency of DMUs with multiple input / output variables. But there are a couple of potential problems by using DEA: (1) Not Suitable for Input / Output Variables with Imprecise Data; (2) DEA requires big data set so data collection is a big problem during assessing DWH efficiency. In general, based on weight computation considerations, if there are m outputs and n inputs we would expect the order of mXn efficient units, suggesting that the number of units in the set should be substantially greater than mXn, in order for there to be suitable discrimination between the units.

Fuzzy Data Envelopment Analysis (FDEA)

One important limitation involves the sensitivity of DEA to the data. Because DEA is a methodology focused on frontiers or boundaries, either noise or errors from data measurement can cause significant problems (Lertworasirikul et al. 2003; Guo and Tanaka 2001; Leon et al 2003). In reality, inputs and outputs are volatile and complex, and cannot satisfy the data accuracy requirement of DEA. Also it is hard to figure out the probability distributions of temporal data, because of lack of enough time-series data. Introducing Fuzzy Logic into DEA will resolve those issues. A Fuzzy number is defined as "a fuzzy set on the real number, which represents the imprecise information such as 'about m'" (Zadeh 1991).

Leon et al 2003 used fuzzy linear programming to allow the input / output data to be interval data instead of traditional DEA using crisp linear programming for the fixed data. This paper fuzzified input / output data by introducing a membership function, then computed a DMU's efficiency based on different possibility levels, ranging from 0, 0.1, 0.2,..,0.9, 1, finally used α-cut, which interacts membership function curve with a cut value, to transform fuzzy linear programming into crisp linear programming so that it can be resolved by standard DEA software. Lertworasirikul et al. 2003 raised an example with two input variables and two output variables, and built up trapezoidal membership functions for those four variables. Once the decision maker specifies the required acceptable (possibility) levels, this fuzzy DEA model can be implemented as easily as crisp DEA models.

FDEA can efficiently process the imprecise input / output data, and adapt to the stochastic change of data set. Lertworasirikul et al. 2003 compared the results from FDEA and DEA and found that FDEA could distinguish the efficiency for some DMUs that standard DEA evaluated at the same efficiency.

FDEA can be formulated as (Lertworasirikul et al. 2003 and Leon et al 2003):

$$E_r = \max \frac{\sum_{i=1}^{m} u_i \tilde{y}_{ri}}{\sum_{i=1}^{n} v_i \tilde{x}_{ri}} \quad \text{Subject to: } \frac{\sum_{i=1}^{m} u_i \tilde{y}_{ri}}{\sum_{i=1}^{n} v_i \tilde{x}_{ri}} \le 1 \quad \text{(For all DMUs, } r=1,2,\ldots s\text{)}$$

Where $\tilde{x}_{ri}$ and $\tilde{y}_{ri}$ are fuzzy number with range.

The mathematical model format of FDEA is very close to that of DEA. The difference is that data for the input and output variables of FDEA are range values, but those of DEA are crisp values. DEA can be viewed as a special version of FDEA using the median value of FDEA.

Solution for FDEA

There have been several FDEA research papers to present different solutions for FDEA problems (Lertworasirikul et al. 2003; Leon et al 2003; Kao and Liu 2000) since the year 2000. In this paper, we adopt the FDEA solution proposed by Kao and Liu 2000, which transferred Fuzzy Linear Programming problems to Linear Programming problems by a set of α-cut values, to assess DWH efficiency. The reason is that the solution offered by Kao and Liu 2000 can convert FDEA problem into traditional DEA problem, and then solve it with the available standard DEA software.

The detailed steps proposed by Kao and Liu 2000 are as follows:

Step 1: Send questionnaires to have DWH users estimate the possible range of input and output variables with uncertainty.

Step 2: Based on the range data from step 1, fuzzify all uncertain data with Triangular Membership Function as Figure 2. The reason for using Triangular Membership Function is that there is a lack of detailed distribution information for the uncertain variables' change, i.e., the estimation range can be regarded as the scope of the fuzzy number with membership value greater than 0, and the mean value as the point with membership value equal to 1. This fuzzifying process applies to all uncertain datasets for all DMUs.

Step 3: Make α-cut for Triangular Membership Function. We make α-cut at 0, 0.25, 0.5, 0.75, and 1.

Step 4: Compute Worst-Best Scenarios.

Calculate the smallest relative efficiency of a DMU compared with other DMUs, i.e., this DMU will be set as "the output level to its smallest value, and the input level to its highest value", and other DMUs will be set as "the output level to its highest value, and the input level to their smallest values".

This DMU (Worst) : output->min, input->max, then efficiency ->min

Other DMUs (Best) : output->max, input->min, then efficiency ->max

$$(E_r)_\alpha^L = \frac{\sum_{i=1}^{m} u_i (y_{ri})_\alpha^L}{\sum_{i=1}^{n} v_i (x_{ri})_\alpha^U} \text{ Subject to: } \frac{\sum_{i=1}^{m} u_i (y_{ri})_\alpha^L}{\sum_{i=1}^{n} v_i (x_{ri})_\alpha^U} \le 1 \text{ and } \frac{\sum_{i=1}^{m} u_i (y_{ji})_\alpha^U}{\sum_{i=1}^{n} v_i (x_{ji})_\alpha^L} \le 1 \ (j \neq r)$$

Where $(E_r)_\alpha^L$ is efficiency of unit *r* at a certain α-cut level for lower bound; $(y_{ri})_\alpha^L$ is amount of output *i* from unit *r* at a certain α-cut level for lower bound; $(y_{ri})_\alpha^U$ is amount of output *i* from unit *r* at a certain α-cut level for upper bound; $(x_{ri})_\alpha^L$ is amount of input *i* from unit *r* at a certain α-cut level for lower bound; $(x_{ri})_\alpha^U$ is amount of input *i* from unit *r* at a certain α-cut level for upper bound.

Step 5: Compute Best-Worst Scenarios.

Calculate the highest relative efficiency of a DMU compared with other DMUs, i.e., this DMU will be set as "the output level to its highest value, and the input level to its smallest value", and other DMUs will be set as "the output level to its smallest value, and the input level to their highest values".

This DMU (Best) : output->max, input->min, then efficiency ->max

Other DMUs: (Worst): output->min, input->max, then efficiency ->min

$$(E_r)_\alpha^U = \frac{\sum_{i=1}^{m} u_i (y_{ri})_\alpha^U}{\sum_{i=1}^{n} v_i (x_{ri})_\alpha^L} \text{ Subject to: } \frac{\sum_{i=1}^{m} u_i (y_{ri})_\alpha^U}{\sum_{i=1}^{n} v_i (x_{ri})_\alpha^L} \le 1 \text{ and } \frac{\sum_{i=1}^{m} u_i (y_{ji})_\alpha^L}{\sum_{i=1}^{n} v_i (x_{ji})_\alpha^U} \le 1 \ (j \neq r)$$

Where $(E_r)_\alpha^U$ is efficiency of unit *r* at certain α-cut level for the upper bound.

Step 6: Defuzzify the fuzzy results, and compute the final fuzzy ranking number.

Regarding the defuzzify algorithm, we use the formula developed by Chen and Klein 1997 to calculate the following index for ranking fuzzy numbers:

$$I_j = \frac{\sum_{i=0}^{t} ((E_j)_{\alpha_j}^U - \min)}{\sum_{i=0}^{t} ((E_j)_{\alpha_j}^U - \min) - \sum_{i=0}^{t} ((E_j)_{\alpha_j}^L - \max)}$$

Where min= $\min_{i,j}\{(E_{ji}\}_{\alpha_i}^L\}$, max= $\max_{i,j}\{(E_{ji}\}_{\alpha_i}^U\}$, *t* is the number of α-cut.

Step 7: Use the results of Step 5 to rank all DWH units.

Step 1~6 presents a way to transform a FDEA problem to a DEA problem and solve it with standard DEA software.

RESEARCH RESULTS AND ANALYSIS

Project Design Considerations

We generated a preliminary data set to gain insights into DEA analysis methods. Based on interviews conducted in (Mannino and Walter, 2003), an initial set of 12 observations was created for the refresh efficiency model. We augmented these 12 observations with 38 additional observations that were randomly generated from a multivariate Gaussian distribution with a mean and covariance matrix based on the original 12 observations. So, the total number of analytic DWHs is 50. Each DWH is regarded as one DMU. Because of the limited data, we only analyze the micro model of refresh efficiency (Figure 2).

The point values were augmented with randomly generated intervals to create fuzzy numbers. For each observation, we generated a random uncertainty level (1: low, 2: moderate, 3: high). For low uncertainty observations, we generated a uniform random number between 0 and 0.10 for each variable to indicate the width of the interval. For moderate and high uncertainty, the widths vary between 0 and 0.25 and 0 and 0.50, respectively.

FDEA Worst-Best Scenario and Best-Worst Scenario

From Table 4 and 5, we only list the first 5 of 50 DWHs' Worst-Best and Best-Worst scenario by use of FDEA (VRS and CRS).

(1) Under Worst-Best Scenario, for every DMU, with the increment of α-cut value, the efficiency of DWH increases. The reason is that the larger α-cut value, the larger the output value and smaller the input value, so the higher the efficiency;

(2) Under Best-Worst Scenario, for every DMU, with the increment of α-cut value, the efficiency of DWH decreases. The reason is that the larger α-cut value, the smaller the output value and larger the input value, so the lower the efficiency.

	Worst-Best					Best-Worst				
DMU	α=0	α=0.25	α=0.5	α=0.75	α=1	α=0	α=0.25	α=0.5	α=0.75	α=1
1	0.348	0.387	0.43	0.485	0.558	1	0.845	0.709	0.633	0.558
2	0.393	0.425	0.457	0.49	0.528	1	1	0.927	0.628	0.528
3	1	1	1	1	1	1	1	1	1	1
4	0.28	0.31	0.349	0.389	0.43	0.91	0.763	0.649	0.493	0.43
5	0.175	0.196	0.222	0.249	0.285	0.503	0.432	0.38	0.331	0.285

Table 4: FDEA (VRS) Results for 50 DWH (only display the first 5 DMUs here)

	Worst-Best					Best-Worst				
DMU	α=0	α=0.25	α=0.5	α=0.75	α=1	α=0	α=0.25	α=0.5	α=0.75	α=1
1	0.3	0.352	0.413	0.472	0.536	0.922	0.799	0.694	0.606	0.536
2	0.242	0.291	0.355	0.429	0.516	1	0.904	0.756	0.627	0.516
3	0.691	0.793	0.912	1	1	1	1	1	1	1
4	0.209	0.254	0.294	0.351	0.412	0.781	0.664	0.575	0.485	0.412
5	0.137	0.165	0.198	0.233	0.273	0.494	0.43	0.365	0.317	0.273

Table 5: FDEA (CRS) Results for 50 DWH (only display the first 5 DMUs here)

Compare Results from FDEA VRS, FDEA CRS, DEA VRS and DEA CRS

To learn the difference in assessment results from DEA and FDEA, we compute the defuzzified values of FDEA, and also the efficiencies based on mean values by standard DEA model. The results are listed at Table 6 and Figure 4.

(1) The overall changing trend of efficiency for DMUs is close between the results of FDEA (VRS or CRS) and those of DEA (VRS or CRS);

(2) Comparing the efficiencies of DEA (VRS) and DEA (CRS) for the same DMU, most of them have close values. But there are slightly different, for DMU 18, 25, 27 and 34, DEA (VRS) ranks them as efficient DMUs, but DEA (CRS) ranks them as inefficient DMUs;

(3) Comparing the efficiencies of FDEA (VRS) and FDEA (CRS) for the same DMU, most of them have close values. Only for DMU 3 and 41, FDEA (VRS) ranks them as efficient DMUs, but FDEA (CRS) ranks them as inefficient DMUs;

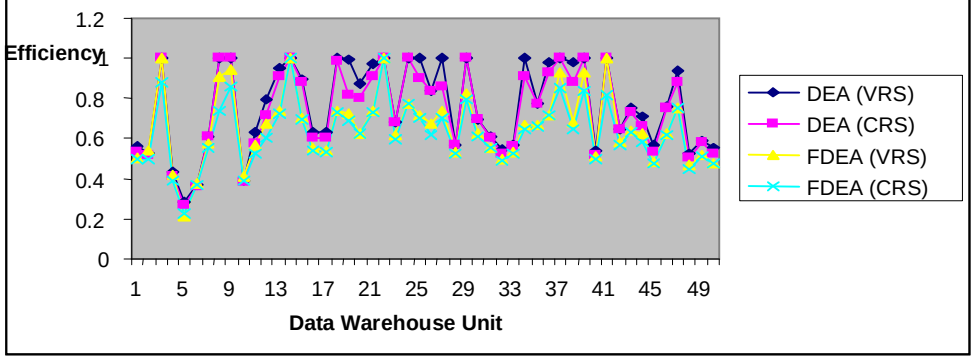
(4) DEA (VRS), DEA (CRS), FDEA (VRS) and FDEA (CRS) rank the efficiency of DMU 14 and 22 are 1, which is the highest efficiency. They also rank DMU 5 as the lowest efficiency one, and DMU 6 as the second lowest efficiency one;

(5) Compared with DEA, FDEA can distinguish the efficiency difference for some DMUs. For example, DEA (VRS and CRS) ranks DMU 8, 9, 24, 29, 37, 39 as efficiency 1, but FDEA (VRS and CRS) can distinguish their efficiencies.

DMU	Final Ranking			
	DEA (VRS)	DEA (CRS)	FDEA (VRS)	FDEA (CRS)
1	0.558	0.536	0.5069	0.4953
2	0.528	0.516	0.5423	0.4961
3	1	1	1	0.8772
4	0.43	0.412	0.4223	0.3908
5	0.285	0.273	0.2142	0.2301

Table 6: Results from DEA (VRS), DEA (CRS), FDEA (VRS) and FDEA (CRS) (only display the first 5 DMUs here)

Figure 4: Results from DEA (VRS, CRS) and FDEA (VRS, CRS)



Group DWHs Based on Efficiency

Table 7 summarizes efficiency rankings for each analysis method in which efficiency scores are grouped into five levels. As depicted in Table 7, scale inefficiencies played a minor role as the DEA columns (DEA CRS and DEA VRS) and the Fuzzy DEA (FDEA) columns (FDEA CRS and FDEA VRS) have minor differences. However, fuzzy DEA had a significant effect on the rankings as shown by comparing the DEA to FDEA columns (DEA CRS to FDEA CRS and DEA VRS to FDEA VRS).

Overall, the more detailed fuzzy DEA analysis showed more pessimistic efficiency evaluations. FDEA also further distinguishes DWHs, such as FDEA CRS and FDEA VRS rank fewer DWHs to be efficient (=1) than DEA CRS and DEA VRS.

Ranking	DEA CRS	DEA VRS	FDEA CRS	FDEA VRS
Efficient (= 1)	14	10	4	2
Near Efficient (> 0.90)	6	5	4	0
Somewhat inefficient (0.75 to 0.90)	7	10	3	7
Inefficient (0.50 to 0.75)	19	21	31	30
Very inefficient (< .50)	4	4	8	11

Table 7: Count of Efficiency Rankings by Analysis Methods

LIMITATIONS AND FUTURE WORKS

The first limitation is that this paper only analyzes the efficiency of the refresh efficiency model, which is one of three DWH efficiency models (Figure 1-3). In the future, we can extend the research to the other two models to complete DWH efficiency research. The second limitation is that we use the hypothetical data set generated by random generator, and we can distribute the survey questionnaire to those companies who have DWH products.

MANAGERIAL IMPLICATIONS

FDEA provides the managers with the ranking of their DWHs, and they can know what efficiency categories their DWHs belong to. Also FDEA offers different choices to evaluate DMU efficiency. The results from Worst-Best scenario give a pessimistic viewpoint for DMUs, which stand for the lowest efficiency for DMUs. The results from Best-Worst scenario give an optimistic viewpoint for DMUs, which stand for the highest efficiency for DMUs.

CONCLUSIONS

We presented efficiency models for evaluating DWH performance and evaluated the relative efficiencies of a preliminary set of DWHs. The efficiency models support evaluation of refresh processing and query production efficiency for a collection of DWHs and individual DWHs. The variables in the models include traditional resource, system usage, data quality, and size measures.

FDEA can determine the efficiencies of DWHs with multiple inputs / outputs, and also handle with imprecise data. The overall changing trend of efficiency for DMUs is close between the results of FDEA (VRS or CRS) and those of DEA (VRS or CRS). The analysis revealed that fuzzy DEA provides more pessimistic efficiency scores, and efficient DWHs processed data from significantly more data sources than inefficient warehouses.

The final ranking results between FDEA and DEA are similar on the highest efficiency and the lowest efficiency of DWHs, but the advantage of FDEA is that it can more deeply distinguish the efficiency difference for some DWHs than standard DEA.

REFERENCES

1. Banker, R.D., Kauffman, R.J. and Morey, R.C. (1990) Measuring Gains in Operational Efficiency from Information Technology: A Study of Positran Deployment at Hardee's Inc. *Journal of Management Information Systems*, 7(2), 29-54.
2. Bouzeghoub, M., Fabret, F. and Matulovic-Broque, M. (1999) Modeling Data Warehouse Refreshment Process as a Workflow Application, *Proceeedings of the International Workshop on Design and Management of Data Warehous*e, Heidelberg, Germany.
3. Charnes, W and Rhodes, E. (1978) Measuring the efficiency of decision making units. *European Journal of Operation. Research*, 2, 429-444.
4. Chen, S. and Klein, C. (1997) A simple approach to ranking a group of aggregated fuzzy utilities, *IEEE Transactions on Systems, Man. Cybernetics - Part B: Cybernetics*, 27(1), 26-35.
5. Delone, W. and McLean, E. (2003) The Delone and McLean Model of Information Systems Success: A Ten-Year Update. *JMIS* 19(4), 9-30.
6. Guo, P. and Tanaka, H. (2001) Fuzzy DEA: a perceptual evaluation method. *Fuzzy Sets & Systems*, 119(2), 149-160.
7. Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P., Vassiliadis, P. (2000). *Fundamentals of Data Warehouses.* Springer Verlag.
8. Kao, C and Liu, S. (2000) Fuzzy Efficiency Measures in Data Envelopment Analysis. *Fuzzy Sets & Systems*, 113(2), 427-437.
9. Leon, T., Liern, V., Ruiz, J. and Sirvent, I. (2003) A fuzzy mathematical programming approach to the assessment of efficiency with DEA models. *Fuzzy Sets & Systems*, 139(2), 407-420.
10. Lertworasirikul, S., Fang, S.C., Joines, J.A., and Nuttle, H.L.W. (2003) Fuzzy data envelopment analysis (DEA): a possibility approach. *Fuzzy Sets & Systems*, 139(2), 379-395.
11. Lin, W. and Shao, B. (2000) Relative Sizes of Information Technology Investments and Productive Efficiency: Their Linkage and Empirical Evidence, *Journal of the AIS* , 1.
12. Mannino, M., Zhang, L. and Choi, I. (2004) *Efficiency Models for Data Warehouse Operations (Working Paper),* Business School, University of Colorado at Denver.
13. Mannino, M. and Walter, Z. (2003) A Field and framework about Data Warehouse Refresh Policies, *Proceedings of the 13th Annual Workshop on Information Technologies & Systems*, Seattle.
14. Orr, K. (1998) Data Quality and Systems Theory. *Communications of the ACM*, 41(2), pp. 66-71.

15. Paradi, J.C., Reese, D.N., and Rosen, D. (1997) Applications of DEA to measure efficiency of software production at two large Canadian banks, *Annals of Operations Research*, 73, 91-115.

16. Pipino, L., Yang, W. L., and Wang, R.Y. (2002) Data Quality Assessment. *Communications of the ACM.*, 45(4), 211-218

17. Shafer, S. and Byrd, T. (2000) A Framework for Measuring the Efficiency of Organizational Investments in Information Technology using Data Envelopment Analysis, *Omega* 28, 125-141.

18. Shao, B. and Shu, W. (2003) Productivity Breakdown of the Information Technology Industries across Countries," in *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS03)*, 238-248.

19. Shin, B. (2003) *An* Exploratory Investigation of System Success Factors in Data Warehousing, *Journal of the Association for Information Systems*, 4, 141-170.

20. Wang, C., Gopal, R., and Zionts, S. (1997) Use of Data Envelopment Analysis in Assessing Information Technology Impact on Firm Performance, *Annals of Operation Research* 73, 191-213.

21. Wixom, B. and Watson, H. (2001) An Empirical Investigation of the Factors Affecting Data Warehousing Success, *MIS Quarterly* 28(1), 17-41.

22. Zadeh, L. A. (1991) *Fuzzy Set Theory and Its Applications*, 2nd ed. Kluwer-Nijhoff, Boston.